GROUP 8

'SPEECH COMMANDS: A DATASET FOR LIMITED-VOCABULARY SPEECH RECOGNITION' BY PETE WARDEN

Prepared by:

- Khairol Izzul Firdaus Bin Khairol Hisam – A21EC0036
- Adam Azhar Bin Nor Adha – A21EC8010
- Muhammad Muadz bin Jamain – B22EC0032
- Muhamad Adib Wafi Bin Muhamad Jais – A21EC0056

VIDEO LINK:
HTTPS://DRIVE.GOOGLE.COM/DRIVE/FOLDERS/1EFAPSHMIXVSNTTLWOE8XQZWOZ3ZKHIXK?USP=SHARING

INTRODUCTION OF THE ARTICLE

- We use speech recognition every day — e.g. “Hey Google”
- Most systems are built for full sentences
- But smart devices often just need to recognize short commands like “yes” or “no”
- Pete Warden created the Speech Commands Dataset for this purpose

PROBLEM STATEMENT OF THE ARTICLE

- Most existing models:
  - Need powerful devices
  - Focus on full sentences
- Real-world use:
  - Simple devices
  - Short commands
  - No good dataset existed for short, one-second commands

SCOPE OF THE ARTICLE

Introduces Speech Commands Dataset:

- **Thousands of 1-second clips**
- **Covers 12+ keywords like “yes”, “no”, “stop”**
- **Collected in noisy and real environments**
- **Goal: Enable lightweight, low-power models for devices like mobile phones**

DATASET PREPARATION

Steps Involved

- Filter out corrupt or short files
- Trim and normalize audio
- Label each clip manually
- Organize into folders (e.g., /yes/, /no/)
- Include noise clips (pink/white noise, background talking)

METHOD & TESTING

- Converted audio into spectrograms (images of sound)
- Used CNN model (like in image recognition)
- Tested with Top-One Accuracy:
- Choose the correct word from 12 options

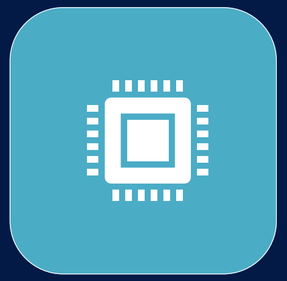
- Result: 88.4% accuracy

STRENGTHS OF THE DATASET

CLEAN AND LABELED DATA

DIVERSE SPEAKERS (2,600+)

EFFECTIVE NOISE HANDLING



EASY FOR DEVELOPERS TO USE

SHORTCOMINGS OF THE ARTICLE

NO REAL-TIME OR STREAMING TESTS

ONLY WORKS ON FIXED COMMANDS

REQUIRES MANUAL SPECTROGRAMS

LACKS NATURAL LANGUAGE UNDERSTANDING

Suggested Improvements

Limitations → Gemini Improvements:

- 1s clips → continuous audio

- Limited vocab → natural language

- Manual spectrograms → raw audio

- No context → intent tracking

Our Working Demo – Overview

- Live prototype with real-time speech recognition

- Web Speech API (client-side speech)

- Socket.IO (real-time communication)

- Gemini API (intelligent interpretation)

- Responsive UI with voice-controlled theme switching

WORKING DEMO – KEY FEATURES

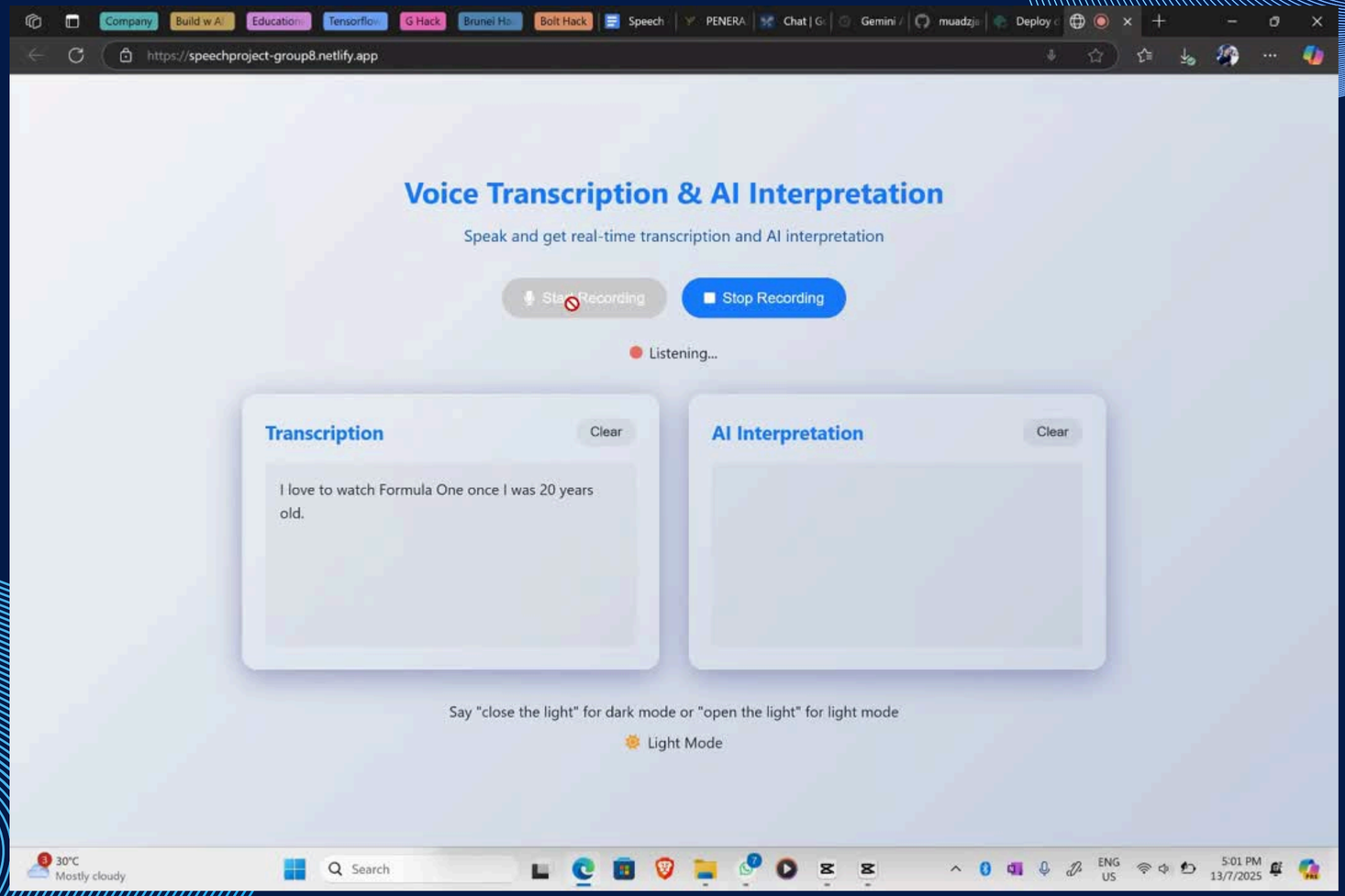
Continuous Listening

Instant Transcription

Gemini Integration

Voice-Activated Control

DEMO


https://speechproject-group8.netlify.app
Voice Transcription & AI Interpretation
Speak and get real-time transcription and AI interpretation
Stop Recording
Listening...
Transcription
Clear
I love to watch Formula One once I was 20 years old.
AI Interpretation
Clear
Say "close the light" for dark mode or "open the light" for light mode
Light Mode
30°C
Mostly cloudy
Search
ENG
US
5:01 PM
13/7/2025

How Our Demo Improves the Paper

- Goes beyond fixed commands → Understands full speech

- Real-time interaction → Not limited to pre-recorded clips

- Smarter response → Understands context using Gemini

- No spectrograms needed → More efficient and modern

Conclusion

- Speech Commands Dataset is a great base

- Good for keyword detection

- Gemini AI enables context-aware, real-time solutions

THANK YOU

GROUP 8